

Improving basic and translational science by accounting for litter-to-litter variation in animal models

Stanley E. Lazic^{*1}, Laurent Essioux¹

¹ Bioinformatics and Exploratory Data Analysis, F. Hoffmann-La Roche, Basel, Switzerland

Email: Stanley E. Lazic^{*} - stan.lazic@cantab.net; laurent.essioux@roche.com;

^{*} Corresponding author

Abstract

Background: Two animals from the same litter are often more alike compared with two animals from separate litters. This litter-to-litter variation (i.e. litter effects) can be either naturally occurring or induced by applying an experimental treatment to whole litters rather than to the individual offspring. An example of the latter is the valproic acid (VPA) model of autism, where the disease phenotype in offspring is caused by giving VPA to pregnant females. In this case, the sample size is the number of pregnant females and not the number of offspring derived from them. If such experiments are not appropriately designed and analysed, the results can be severely biased as well as extremely underpowered.

Results: A review of the VPA literature showed that only 9% (3/34) of studies correctly determined the sample size. In addition, litter effects accounted for up to 61% ($p < 0.001$) of the variation in behavioural outcomes, which was much larger than the treatment effects. In addition, few studies reported using randomisation (12%) or blinding (18%), and none indicated that a sample size calculation or power analysis had been conducted.

Conclusions: Litter effects are common, large, and ignoring them can make replication of findings difficult and can contribute to the low rate of translating preclinical in vivo studies into successful therapies. Only a minority of studies reported using rigorous experimental methods, which is consistent with much of the preclinical in vivo literature.

Key words: Autism, Experimental design, Litter-effects, Mixed-effects model, Multiparous, Nested model, Valproic acid

Background

Numerous animal models (lesion, transgenic, knock-out, selective breeding, etc.) have been developed for a variety of psychiatric, neurodegenerative, and neurodevelopmental disorders. While many of these models have been helpful for understanding disease pathology, they have been less useful for discovering potential therapies, or for predicting which

treatments will be useful in the clinic. Translation from in vivo animal models (typically rodent) has been poor, despite many years of research and effort. There are many reasons for this, including the inherent difference in biology between rodents and humans [1], particularly relating to higher cognitive functions. In addition, there is the ever-present question of whether a particular animal model is

even suitable; whether it captures the disease process of interest or faithfully mimics key aspects of the human condition. While important, these two considerations will be put aside and the focus will be on the design and analysis of a key aspect of pre-clinical studies using multiparous species, and the role that this has on the reproducibility of results. There are two issues that will be discussed. The first deals with designs where an experimental treatment is applied to whole litters rather than to the individual animals, usually because the treatment is applied to pregnant females and therefore to all of the offspring. The second is the natural litter-to-litter variation (i.e. litter effect) that is often present, which means that the value of a measured experimental outcome is influenced by the litter that the animal came from.

Applying treatments to whole litters

Some disease models have a distinctive experimental design feature: the treatment is applied to pregnant females (and therefore to all of the unborn animals within that female), but the scientific interest is in the individual offspring (Figure 1). Here, the “treatment” refers to the experimental manipulation that induces the disease features, and it does not refer to a therapeutic treatment. This design is common in toxicology and nutrition studies, but also used in neuroscience studies when examining the effects of maternal stress and in the valproic acid (VPA) model of autism. Difficulties arise because the experimental unit (“ n ”; defined as the smallest physical unit that can be randomly assigned to a treatment condition) is the pregnant dam and not the individual offspring [2–11]. In other words, the sample size is the number of dams, and the offspring are considered subsamples, much like the left and right kidney from a single animal do not represent a sample size of two ($n = 1$, but there are two replicate measurements). This may come as a surprise, and it is irrelevant that the scientific interest is in the offspring, or that the offspring eventually become individual entities (unlike kidneys). Regulatory authorities have clear guidelines on the matter [12, 13]; for example, the Organisation for Economic Co-operation and Development (OECD) has made a firm statement in their guidelines for chemical testing: “Developmental studies

using multiparous species where multiple pups per litter are tested should include the litter in the statistical model to guard against an inflated Type I error rates. The statistical unit of measure should be the litter and not the pup. Experiments should be designed such that littermates are not treated as independent observations [p. 12]” [13]. There is a restriction on randomisation because only whole litters can be assigned to the treatment or control conditions, which has implications for how studies are designed and analysed. An appropriate analysis can be conducted by using only one animal per litter (randomly selected), which allows standard methods to be used (e.g. t-test, ANOVA, etc.). This is often not the most efficient design in terms of animal usage, unless the excess animals can be used for other experiments. A second option is to use more than one animal per litter, and then average the values of the animals within a litter. These mean values can then be taken forward and analysed using standard methods. A third option is to use multiple animals per litter, and then analysis is performed with a nested or hierarchical model, which properly handles the structure of the data (i.e. animals are nested within litters) and avoids artificially inflating the sample size (also known as pseudoreplication [11, 14]). The third method is preferred over averaging values within a litter because litter is entered as a variable in the analysis and the magnitude of the litter effect can be quantified. In addition, information on the precision will be lost by averaging, but is retained and made use of in the hierarchical model. When using the first two options, it is clear that to increase the sample size and thus power, the number of litters needs to be increased. This is also true for the third option, but may not be so readily apparent [9, pp. 3–4], and is discussed further below.

A related design issue is that greater statistical power can be achieved when littermates are used to test a therapeutic compound versus a placebo. If the therapeutic treatment is applied to the individual animals postnatally, then the individual animal is the experimental unit *for this comparison*. This is referred to as a split-plot design and has more than one type of experimental unit: litters for some comparisons and individual animals for others. These studies therefore require careful planning and analysis, but biologists are rarely introduced to these designs and how to appropriately analyse them

during the course of their training.

Litter effects are ubiquitous, large, and important

It is known that on many variables and across many species, monozygotic twins are more similar than dizygotic twins, which are more similar than non-twin siblings, and which in turn are more similar than two unrelated individuals. What has not been fully appreciated is that all of the standard statistical methods (e.g. t-test, ANOVA, regression, non-parametric methods) assume that the data come from unrelated individuals. However, rodents from the same litter are effectively dizygotic twins; they are genetically very similar and share prenatal and early postnatal environments. Therefore studies need to be designed and analysed in such a way that differences between litters do not bias or confound the results [2–11]. More specifically, this relates to the assumption of independence of observations. For example, measuring blood pressure (BP) from the left and right arm of ten unrelated people only provides ten independent measurements of BP, not twenty. This is because the left and right BP values will be highly correlated—if the BP value measured from a person’s left arm is high, then so will the value measured from their right arm. Similarly, two animals from the same litter will tend to have values that are more alike (i.e. correlated) than two animals from two different litters. This lack of independence needs to be handled appropriately in the analysis and the three strategies outlined in the previous section can be used. Many animal models are derived from highly inbred strains, and this results in reduced genotypic and phenotypic variation. This is a different issue and unrelated to lack of independence. It does not mean that animals “are all the same” and that differences between litters do not exist.

Litter effects are not a minor issue that only statistical pedants worry about, with little practical importance for scientists. Using actual body weight data from their experiment Holson and Pearce showed twenty years ago that if three treated and three control litters are used, with two offspring per litter (total number of offspring = 12), then the false positive rate (Type I error) is 20% rather

than 5% [3]. Furthermore, the false positive rate increases with the number of offspring per litter: if the number of offspring per litter is 12 (total number of offspring = 72) then the false positive rate is 80%. The error rate is also influenced by the relative variability between and within litters and will therefore vary for each experimental outcome. Nevertheless, given that papers report the results of multiple tests (multiple outcome variables and multiple comparisons), we can expect the literature to be rife with false positive results. It may seem paradoxical, but in addition to too many false positives, ignoring litter-to-litter variation can also lead to low power (too many false negatives) when true effects exist [3, 4]. This is because differences between litters ends up as unexplained variation, and thus the “noise” in the data is increased, potentially masking true treatment effects. A subsequent study in 1997 using forty litters found “significant litter effects... in varying degrees, for almost every behavioural, morphologic, and neuroendocrine measure; they were evident across indices of neural, adrenal, thyroid, and immunologic functioning in adulthood” [4] (and see references therein for further studies supporting this conclusion). Holson and Pearce reported that only 30% of papers in the behavioural neurotoxicology literature correctly accounted for litter effects [3] and Zorrilla noted that 34% of papers in *Developmental Psychobiology* correctly accounted for litter effects and only 15% of papers in related journals [4]. This issue has been brought up repeatedly for almost forty years [2], but has largely been ignored by experimental biologists. One can only speculate on the number of erroneous conclusions that have been reached, and the resources that have been wasted.

One might argue that when many studies are conducted, including replications within and between labs, the evidence will eventually converge to the “truth”, and therefore these considerations are only of minor interest. Unfortunately, there is no guarantee of such convergence, as the literature on the superoxide dismutase (SOD1) transgenic mouse model of amyotrophic lateral sclerosis (ALS) demonstrates. Several treatments showed efficacy in this model and were advanced to clinical trials, where they proved to be ineffective [15]. A subsequent large-scale and properly executed replication study did not support the previous findings [16]. This

study also identified litter as an important variable which affected survival (the main outcome), and which was not taken into account in the earlier studies. The authors also demonstrated how false positive results can arise with inappropriate experimental designs and analyses. Litter effects were not the only contributing factor; a meta-analysis of the preclinical SOD1 literature revealed that only 31% of studies reported randomly assigning animals to treatment conditions, and even fewer reported blind assessment of outcomes [17]. Lack of randomisation and blinding are known to overstate the size of treatment effects [18–22]. In addition, there was evidence of publication bias, where studies with positive results were more likely to be published [17]. Thus, the combination of poor experimental design, analysis, and publication bias contributed to numerous incorrect decisions regarding treatment efficacy.

General quality of preclinical animal studies

Previous studies have shown that general quality of the design, analysis, and interpretation of preclinical animal experiments is low [19, 20, 22–29]. For example, Nieuwenhuis et al. recently reported that 50% of papers in the neuroscience literature misinterpret interaction effects [30]. In addition, the issue of “inflated n ”, or pseudoreplication, shows up in other guises [11, 31], and whole fields can misattribute cause-and-effect relationships [32, 33]. There is also the concept of “researcher degrees of freedom”, which refers to the flexibility that scientists have in choosing the main outcome variables, statistical models, data transformations, how outliers are handled, when to stop collecting data, and what is reported in the final paper [34]. Various permutations of the above options greatly increases the chances that at least something will be statistically significant, and this is what tends to get reported as the sole analysis that was conducted. Given the above concerns, it is not surprising that the pharmaceutical industry has difficulty reproducing many published results [35–38].

Methods

Literature review

Primary research articles that injected pregnant dams with VPA and subsequently analysed the effects in the offspring were identified on PubMed using the search term “(VPA OR ‘valproic acid’) AND autism” (up to the end of 2011). Reference lists from these articles were then examined for further relevant studies. A total of thirty-five studies were found, and one was excluded as key information was located in the supplementary material, but this was not available online [39]. Two key pieces of information were extracted: (1) whether the analysis correctly identified the experimental unit as the litter, and (2) whether important features of good experimental design were mentioned, including randomisation, blinding, sample size calculation, and whether the total sample size (i.e. number of pregnant dams) was indicated or could be determined.

Estimating the importance of litter-to-litter variation

Data from Mehta et al. [40] were used to estimate the magnitude of differences between litters on a number of outcome variables. This study was chosen because it included animals from fourteen litters (five saline, nine VPA) and therefore it was possible to get a good estimate of the litter-to-litter variation. In addition, the study mentioned using randomisation and blind assessment of outcomes. Half of the animals in each condition were also given MPEP (2-methyl-6-phenylethyl-pyrididine), a metabotropic glutamate receptor 5 antagonist. To assess the magnitude of the litter effects, the effect of VPA, MPEP, and sex (if relevant) were removed, and the remaining variability in the data that could be attributed to differences between litters was estimated. More specifically, models with and without a random effect of litter were compared with a likelihood ratio test. This analysis is testing whether the variance between litters is zero, and it is known that p-values will be too large because of “testing on the boundary”, and so the simple method of dividing the resulting p-values by two was used as recommended by Zuur et al [41]. The exact specification of the models is provided as R code in Additional

File 1 and the data are provided in Additional File 2.

Power analysis

In these types of designs, power (the ability to detect an effect that is actually present) is influenced by (1) the number of litters, (2) variability between litters, (3) number of animals within litters, (4) variability of animals within litters, (5) difference between the means of the treatment groups (effect size), (6) significance cutoff (traditionally $\alpha = 0.05$), and (7) the statistical test used. In order to illustrate the importance of the number of litters relative to the number of animals within litters, a power analysis was conducted with the number of litters per group varying from three to ten, and the number of animals per litter varying from one to ten. The other factors were held constant. Variability between litters ($SD = 0.8803$) and the variability of animals within litters ($SD = 0.8142$) was estimated from the locomotor activity data from Mehta et al. [40]. For each combination of litters and animals, 5000 simulated datasets were created with a mean difference between groups of 0.15. Once the datasets were generated, the power for three types of analyses were calculated. The first analysis averaged the values of the animals within each litter, and then groups were compared with a t-test. The second analysis used a mixed-effects model, and the third ignored litter and just compared all of the values groups with a t-test. The last analysis is incorrect and only presented to demonstrate how artificially inflating sample size affects power. The power for each analysis was determined as the proportion of tests that had $p < 0.05$. The R code is provided in Additional File 1 and is adapted from Gelman and Hill [42].

Results and Discussion

Low quality of the published literature

The VPA model of autism is relatively new and potential therapeutic compounds tested in this model have not yet advanced to human trials. The opportunity therefore exists to clean up the literature and prevent a repeat of the SOD1 story. The main finding is that only 9% (3/34) of studies correctly

identified the experimental unit and thus made valid inferences from the data. One study used a nested design [43], the second mentioned that litter was the experimental unit [44], and the third used one animal from each litter, thus bypassing the issues discussed [45]. For fourteen studies (41%) it was not possible to determine how many dams were actually used (i.e. the sample size), and in four studies (12%) the number of offspring used were not indicated. In addition, only four (12%) reported randomly assigning pregnant females to the VPA or control group. Many studies also used only a subset of the offspring from each litter, but often it was not mentioned how the offspring were selected. Only six studies (18%) reported that the investigator was blind to the experimental condition when collecting the data. Ten studies (29%) did not indicate whether both male and female offspring were used. No study mentioned performing a power analysis to determine a suitable sample size to detect effects of a given magnitude—but this is probably fortuitous, given that only three studies correctly identified the experimental unit. It is possible that many studies did actually randomise and assess outcomes blind but simply did not report it. However, randomisation and blinding are crucial aspects for the validity of the results and their omission in manuscripts suggests that they were not used. This is further supported by studies showing that when manuscripts do not mention using randomisation or blinding, the estimated effects sizes are larger compared to studies that do mention using these methods, which indicative of bias [18–22, 28].

A number of papers had additional statistical or experimental design issues, ranging from trivial (e.g. reporting total degrees of freedom rather than residual degrees of freedom for an F-statistic) to serious. These include treating individual neurons as the experimental unit, which is distressingly common in electrophysiological studies but just as inappropriate as treating blood pressure values taken from left and right arms as $n = 2$, or chopping a single liver sample into ten pieces and treating the expression of a gene measured in each piece as $n = 10$ [11]. If only it were so easy; clinical trials could be conducted with tens of patients rather than hundreds or thousands. Regulatory authorities are not fooled by such stratagems, but it seems many journal editors and

peer-reviewers are.

In addition, the wording of two studies suggests that control dams did not receive a vehicle injection, and thus any differences between groups may be partly due to the stress of handling and injection [46, 47]. For some studies, the reported degrees of freedom did not correspond to what would be expected based on the verbal description of the analysis, and a number of studies did not correctly distinguish between “within-subjects” and “between-subjects” effects. A list of studies can be found in Additional File 3.

Estimating the magnitude of litter effects

To estimate the extent to which litter effects are important and how they can affect the results, data originally published by Mehta et al. [40] were used, and experimental details can be found therein. Locomotor activity in the open field is shown in Figure 2 for nine VPA and five saline injected controls litters. Half of the animals from each condition were given MPEP (a mGluR5 receptor antagonist) or saline. There do not appear to be differences between VPA and control groups, and a slight increase in activity due to MPEP. This effect of MPEP was not significant when litter effects were ignored (Figure 2A; $p = 0.082$), but it was when adjusting for litter (Figure 2B; $p = 0.011$). In this case the shift in p-value was not large, but it happened to decrease it below the 0.05 threshold after the excess noise caused by litter-to-litter variation was removed.

It may be difficult to determine whether litter effects are present by simply plotting the data by litter because the experimental factors—especially if they are large—may obscure the effects. It is therefore better to remove the effect of the experimental factors first, and then plot the residual values versus litter. The y -axis for Figure 3 plots the residuals, which is the difference between the observed locomotor activity for each animal and the value predicted from a model containing group (VPA/saline) and condition (MPEP/saline) as factors (from Figure 2A). The residuals should be pure noise, centred at zero, and should not be associated with any other

variable. However, it is clear that there are large differences between litters (Figure 3A), indicating heterogeneity in the response from one litter to the next. When litter effects are taken into account, the mean of each litter is closer to zero. Also note that variance of the residuals (σ_ϵ^2) is reduced by 61% when litter is taken into account ($p < 0.001$). This is shown by the spread of the grey points around zero on the right side of each graph, which are clustered closer together in the second analysis. This means that litter accounted for 61% of the previously unexplained variation in the data. Note that it would be impossible to determine whether litter effects are present if only one litter per treatment group was used because litter and treatment would be completely confounded.

A similar analyses was performed for other variables and the results are displayed in Table 1. It is clear that litter-to-litter variation is important for a number of behavioural outcomes. It is also clear from Figure 3A how easy it is to get false positives with an inappropriate design and analysis. Imagine if an experiment was conducted with only one VPA and one saline litter, with ten animals from each, and that there is no overall effect of VPA on a particular outcome. If the experimenter happened to select Litter A (saline) and Litter M (VPA) there would be a significant increase due to VPA, but if Litter D (saline) and Litter G (VPA) were selected, there would be a significant effect in the opposite direction! There are many combinations of a single saline and VPA litter that would lead to a significant difference between conditions. Having two or three litters per group instead of one will reduce the false positive rate, but it will still be much higher than 0.05 [3]. In addition, these apparent differences would not replicate with a properly designed experiment.

How power is affected by the number of litters and animals

Figure 4 shows the power for various combinations of number of litters and number of animals per litter. This analysis is based on averaging the values for the animals within a litter and then comparing the groups with a t-test. It is clear that increasing

the number of animals per litter has little effect on power (the lines in Figure 4A are nearly flat after two animals per litter), whereas increasing the number of litters results in a large increase power. The results for the mixed-effect model are nearly identical, and the results of the inappropriate analysis which ignores litter shows increasing power with increasing number of animals per litter (Additional File 4). This is false power however, and is due to an artificially inflated sample size (pseudoreplication), and will lead many false positive results.

Some may object on ethical grounds to using so many litters and then taking only one or a few animals from each, as there will be many additional animals that will not be used, and presumably culled. Certainly all of the animals could be used, but there is almost no increase in power after three animals per litter (at least for the locomotor data) and therefore it is a poor use of time and resources to include all of the animals. One could argue therefore that it is unethical to submit a greater number of animals to the experimental procedure if they contribute little or nothing to the result. One could also argue that it is even more unethical to use any animals for a flawed or severely underpowered study in the first place, and then to clutter the scientific literature with the results. One way to deal with this issue is to use the excess animals for other experiments. For example, a few animals per litter might be used for a novel behavioural task. Others may be used to test the effects of a therapeutic compound, and rest for a study looking at gene expression. This requires greater planning, organisation, and coordination, but it is possible. Another option is to purchase animals from a supplier and request that the animals come from different litters rather than have an in-house colony.

How does litter-to-litter variation arise?

Differences between litters could exist for a variety of reasons, including shared genes and shared prenatal and early postnatal environments, but also due to age differences (it is difficult to control the time of mating), and because litters are convenient units to work with. For example, it is not unusual for litter-mates to be housed in the same cage, which means

that animals within a litter also share not just their early, but also their adult environment. It is also often administratively easier to apply experimental treatments on a per-cage (and thus per-litter) basis rather than per-animal basis. For example, animals in cage A and C are treated while cage B and D are controls. Animals may also undergo behavioural testing on a per-cage basis; for example, animals are taken from the housing room to the testing room one cage at a time, tested, and then returned. Larger experiments may need to be conducted over several days, and it is often convenient to do four cages on one day and four on the next, rather than take half the animals from all eight cages on each day. At the end of the experiment animals may also be killed on a per-cage basis. Given that it may take many hours to kill the animals, remove brains, collect blood, etc., the values of many outcomes (e.g. gene expression, hormones and metabolites concentration, physiological parameters, etc.) will change due to circadian rhythms. All of these can lead to systematic differences between litters and can thus bias results and/or add noise to the data.

There is an important distinction to be made between applying treatments to whole litters versus “natural” variation between litters. When a treatment is applied to a whole litter (e.g. VPA model of autism, maternal stress) then the litter is the experimental unit and thus the sample size is the number of litters. Therefore, *by definition*, litter needs to be included in the analysis if more than one animal per litter is used (or the values within a litter can be averaged). However, if multiple litters are used but the treatment(s) are applied to the individual animals, experiments should be designed so that *if* litter effects exist, then valid inferences can still be made. In other words, litters should not be confounded with other experimental variables, because it would be difficult or impossible to detect their influence and remove their effects. Whether litter is an important factor for any particular outcome is then an empirical question, and if it is not important then it need not be included in the analysis. However, the power to detect differences between litters will be low if only a few litters are used in the experiment, and therefore a non-significant test for litter effects should not be interpreted as the absence of such effects. What should *not* be done is to analyse the data with and without litter and choose

the analysis that gives the “right” answer for the experimental variable of interest [34]. Flood et al. provide an nice example in the autism literature of an appropriate design followed by an analysis both with and without litter included [48]. They also found a strong effect of litter on brain mass.

Four ways to improve basic and translational research

Better training for biologists

Most experimental biologists are not provided with sufficient training in experimental design and data analysis to be able to plan, conduct, and interpret the results of scientific investigations at the level required to consistently obtain valid results. The solution is straightforward but requires major changes in the education and training of biologists and it will take many years to implement. Nevertheless, this should be a longer-term goal for the biomedical research community.

Make better use of statistical expertise

An second solution is to have statisticians play a greater role in preclinical studies, including peer reviewing grant applications and manuscripts, as well as being part of scientific teams [49]. However, there are not enough statisticians with the appropriate subject matter knowledge to fully meet this demand—just as it is difficult to do good science without a knowledge of statistics, it is difficult to perform a good analysis without knowledge of the science. In addition, this type of “project support” is often viewed by academic statisticians as a secondary activity. Despite this, there is still scope for improving the quality of studies by making better use of statistical expertise.

More detailed reporting of experimental methods

Detailed reporting of how experiments were conducted, how data were analysed, how outliers were handled, whether all animals that entered the study

completed it, and how the sample size was determined are all required to assess whether the results of the study are valid, and a number of guidelines have been proposed which cover these points, including the National Institute of Neurological Disorders and Stroke (NINDS) guidelines [50], the Gold Standard Publication Checklist [51], and the ARRIVE (Animals in Research: Reporting In Vivo Experiments) guidelines [52]. For example, ARRIVE items 6 (Study design), 10 (Sample size), 11 (Allocating animals to experimental groups), and 13 (Statistical methods) should be a mandatory requirement for all publications involving animals and could be included as a separate checklist that is submitted along with the manuscript, much like a conflict of interest or a transfer of copyright form. This would make it easier to spot any design and analysis issues by reviewers, editors, and other readers. In addition, and more importantly, if scientists are *required* to comment on how they randomised treatment allocation, or how they ensured that assessment of outcomes was blinded, then they will conduct their experiments accordingly if they plan on publishing in a journal that has these reporting requirements. Similarly, if researchers are required to state what the experimental unit is (e.g. litter, cage, individual animal, etc.), then they will be prompted to think hard about the issue and design better experiments, or seek advice. This recommendation will not only improve the quality of reporting, but it will also improve the quality of experiments, which is the real benefit. A final benefit is that it will make quantitative reviews/meta-analyses easier, because much of the key information will be on a single page.

Make raw data available

Another solution is to make the provision of raw data a requirement for acceptance of a manuscript; not “to make it available if someone asks for it”, which is the current requirement for many journals, but uploaded as supplementary material or hosted by a third party data repository. None of the VPA studies provided the data that the conclusions were based on, making reanalysis impossible. Remarkably, of the thirty-five studies published, only one provided the necessary information to conduct a power analysis to plan a future study [45], and this was only because one animal per litter was used and

the necessary values could be extracted from the graphs. Datasets used in preclinical animal studies are typically small, do not have confidentiality issues associated with them, are unlikely to be used for further analyses by the original authors, and have no additional intellectual property issues associated with them given that the manuscript itself has been published. It is noteworthy that many journals require microarray data to be uploaded to a publicly available repository (e.g. Gene Expression Omnibus or ArrayExpress), but not the corresponding behavioural or histological data. It is perhaps not surprising that there is a relationship between study quality and the willingness to share data [53–55]. Publishing raw data can be taken as a signal that researchers stand behind their data and therefore their conclusions. Funding bodies should encourage this by requiring that data arising from the grant are made publicly available (with penalties for non-adherence).

The above suggestions would help ensure that appropriate design and analyses were used, and to make it easy to verify claims or to reanalyse data. Currently, it is often difficult to establish the former and almost impossible to perform the latter. Moreover, it is clear that appropriate designs and analyses are often not used, making it difficult to give the benefit of the doubt to those studies with incomplete reporting of how experiments were conducted and data analysed.

Conclusions

While it is difficult to quantify the extent to which poor statistical practices hinder basic and translational research, it is clear that a large inflation of false positive and false negative rates will only slow progress down. In addition, because of publication bias and researcher degrees of freedom, it is possible for a field to converge to the wrong answer. Experimental design and statistical issues are, in principle, fixable. Improving these will allow scientists to focus on creating and assessing the suitability of disease models and the efficacy of therapeutic interventions, which is challenging enough.

List of abbreviations

ANOVA: analysis of variance; BP: blood pressure; MPEP: 2-methyl-6-phenylethyl-pyrididine; SOD1: superoxide dismutase; VPA: valproic acid

Authors contributions

SEL planned and carried out the study, performed the literature search and analysis, and wrote the paper. LE provided constructive input. All authors read and approved the final manuscript.

Competing interests

The author declares no competing interests.

Acknowledgements

The authors would like to thank the Siegel lab at the University of Pennsylvania for kindly sharing their data.

References

1. Geerts H: **Of mice and men: bridging the translational disconnect in CNS drug discovery.** *CNS Drugs* 2009, **23**(11):915–926, [<http://dx.doi.org/10.2165/11310890-000000000-00000>].
2. Haseman JK, Hogan MD: **Selection of the experimental unit in teratology studies.** *Teratology* 1975, **12**(2):165–171.
3. Holson RR, Pearce B: **Principles and pitfalls in the analysis of prenatal treatment effects in multiparous species.** *Neurotoxicol Teratol* 1992, **14**(3):221–228.
4. Zorrilla EP: **Multiparous species present problems (and possibilities) to developmentalists.** *Dev Psychobiol* 1997, **30**(2):141–150.
5. Wainwright PE: **Issues of design and analysis relating to the use of multiparous species in developmental nutritional studies.** *J Nutr* 1998, **128**(3):661–663.
6. Festing MFW, Altman DG: **Guidelines for the design and statistical analysis of experiments using laboratory animals.** *ILAR J* 2002, **43**(4):244–258.
7. Festing MFW: **Principles: the need for better experimental design.** *Trends Pharmacol Sci* 2003, **24**(7):341–345.
8. Festing MFW: **Design and statistical methods in studies using animal models of development.** *ILAR J* 2006, **47**:5–14.
9. Casella G: *Statistical Design*. New York : Springer 2008.

10. Maurissen J: **Practical considerations on the design, execution and analysis of developmental neurotoxicity studies to be published in Neurotoxicology and Teratology.** *Neurotoxicol Teratol* 2010, **32**(2):121–123, [http://dx.doi.org/10.1016/j.ntt.2009.09.002].
11. Lazic SE: **The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis?** *BMC Neurosci* 2010, **11**:5, [http://dx.doi.org/10.1186/1471-2202-11-5].
12. International Conference on Harmonisation: **Detection of toxicity to reproduction for medicinal products and toxicity to male fertility.** *S5(R2)* 1993.
13. OECD: **Guideline for the testing of chemicals: developmental neurotoxicity study** 2007, :1–26, [http://www.oecd-ilibrary.org/test-no-426-developmental-neurotoxicity-study_514fg25mnkxs.pdf;jsessionid=12ic71yg7bopl.delta?contentType=/ns/Book&itemId=/content/book/9789264067394-en&containerItemId=/content/serial/20745788&accessItemIds=&mimeType=application/pdf].
14. Hurlbert SH: **Pseudoreplication and the design of ecological field experiments.** *Ecol Monogr* 1984, **54**(2):187–211.
15. Schnabel J: **Neuroscience: Standard model.** *Nature* 2008, **454**(7205):682–685, [http://dx.doi.org/10.1038/454682a].
16. Scott S, Kranz JE, Cole J, Lincecum JM, Thompson K, Kelly N, Bostrom A, Theodoss J, Al-Nakhala BM, Vieira FG, Ramasubbu J, Heywood JA: **Design, power, and interpretation of studies in the standard murine model of ALS.** *Amyotroph Lateral Scler* 2008, **9**:4–15, [http://dx.doi.org/10.1080/17482960701856300].
17. Benatar M: **Lost in translation: treatment trials in the SOD1 mouse and in human ALS.** *Neurobiol Dis* 2007, **26**:1–13, [http://dx.doi.org/10.1016/j.nbd.2006.12.015].
18. Bebarta V, Luyten D, Heard K: **Emergency medicine animal research: does use of randomization and blinding affect the results?** *Acad Emerg Med* 2003, **10**(6):684–687.
19. Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PMW, Macleod M, Dirnagl U: **Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach.** *Stroke* 2008, **39**(3):929–934, [http://dx.doi.org/10.1161/STROKEAHA.107.498725].
20. Kilkenny C, Parsons N, Kadyszewski E, Festing MFW, Cuthill IC, Fry D, Hutton J, Altman DG: **Survey of the quality of experimental design, statistical analysis and reporting of research using animals.** *PLoS One* 2009, **4**(11):e7824, [http://dx.doi.org/10.1371/journal.pone.0007824].
21. Sena ES, van der Worp HB, Bath PMW, Howells DW, Macleod MR: **Publication bias in reports of animal stroke studies leads to major overstatement of efficacy.** *PLoS Biol* 2010, **8**(3):e1000344, [http://dx.doi.org/10.1371/journal.pbio.1000344].
22. Vesterinen HM, Sena ES, French Constant C, Williams A, Chandran S, Macleod MR: **Improving the translational hit of experimental treatments in multiple sclerosis.** *Mult Scler* 2010, **16**(9):1044–1055, [http://dx.doi.org/10.1177/1352458510379612].
23. Hackam DG, Redelmeier DA: **Translation of research evidence from animals to humans.** *JAMA* 2006, **296**(14):1731–1732, [http://dx.doi.org/10.1001/jama.296.14.1731].
24. Dirnagl U: **Bench to bedside: the quest for quality in experimental stroke research.** *J Cereb Blood Flow Metab* 2006, **26**(12):1465–1478, [http://dx.doi.org/10.1038/sj.jcbfm.9600298].
25. Philip M, Benatar M, Fisher M, Savitz SI: **Methodological quality of animal studies of neuroprotective agents currently in phase II/III acute ischemic stroke trials.** *Stroke* 2009, **40**(2):577–581, [http://dx.doi.org/10.1161/STROKEAHA.108.524330].
26. Bart van der Worp H, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR: **Can animal models of disease reliably inform human studies?** *PLoS Med* 2010, **7**(3):e1000245, [http://dx.doi.org/10.1371/journal.pmed.1000245].
27. Shineman DW, Basi GS, Bizon JL, Colton CA, Greenberg BD, Hollister BA, Lincecum J, Leblanc GG, Lee LBH, Luo F, Morgan D, Morse I, Refolo LM, Riddell DR, Searce-Levie K, Sweeney P, Yrjanheikki J, Fillit HM: **Accelerating drug discovery for Alzheimer's disease: best practices for preclinical animal studies.** *Alzheimers Res Ther* 2011, **3**(5):28, [http://dx.doi.org/10.1186/alzrt90].
28. Rooke EDM, Vesterinen HM, Sena ES, Egan KJ, Macleod MR: **Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis.** *Parkinsonism Relat Disord* 2011, **17**(5):313–320, [http://dx.doi.org/10.1016/j.parkreldis.2011.02.010].
29. Sarewitz D: **Beware the creeping cracks of bias.** *Nature* 2012, **485**(7397):149, [http://dx.doi.org/10.1038/485149a].
30. Nieuwenhuis S, Forstmann BU, Wagenmakers EJ: **Erroneous analyses of interactions in neuroscience: a problem of significance.** *Nat Neurosci* 2011, **14**(9):1105–1107, [http://dx.doi.org/10.1038/nn.2886].
31. Cumming G, Fidler F, Vaux DL: **Error bars in experimental biology.** *J Cell Biol* 2007, **177**:7–11.
32. Lazic SE: **Relating hippocampal neurogenesis to behavior: the dangers of ignoring confounding variables.** *Neurobiol Aging* 2010, **31**(12):2169–2171, [http://dx.doi.org/10.1016/j.neurobiolaging.2010.04.037].
33. Lazic SE: **Using causal models to distinguish between neurogenesis-dependent and -independent effects on behaviour.** *J R Soc Interface* 2011; e-pub ahead of print 28 Sep 2011; doi:10.1098/rsif.2011.0510, [http://dx.doi.org/10.1098/rsif.2011.0510].
34. Simmons JP, Nelson LD, Simonsohn U: **False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant.** *Psychol Sci* 2011, **22**(11):1359–1366, [http://dx.doi.org/10.1177/0956797611417632].

35. Mullard A: **Reliability of 'new drug target' claims called into question.** *Nat Rev Drug Discov* 2011, **10**(9):643–644, [http://dx.doi.org/10.1038/nrd3545].
36. Prinz F, Schlange T, Asadullah K: **Believe it or not: how much can we rely on published data on potential drug targets?** *Nat Rev Drug Discov* 2011, **10**(9):712, [http://dx.doi.org/10.1038/nrd3439-c1].
37. Ledford H: **Drug candidates derailed in case of mistaken identity.** *Nature* 2012, **483**(7391):519, [http://dx.doi.org/10.1038/483519a].
38. Begley CG, Ellis LM: **Drug development: Raise standards for preclinical cancer research.** *Nature* 2012, **483**(7391):531–533, [http://dx.doi.org/10.1038/483531a].
39. Rinaldi T, Silberberg G, Markram H: **Hyperconnectivity of local neocortical microcircuitry induced by prenatal exposure to valproic acid.** *Cereb Cortex* 2008, **18**(4):763–770, [http://dx.doi.org/10.1093/cercor/bhm117].
40. Mehta MV, Gandal MJ, Siegel SJ: **mGluR5-antagonist mediated reversal of elevated stereotyped, repetitive behaviors in the VPA model of autism.** *PLoS One* 2011, **6**(10):e26077, [http://dx.doi.org/10.1371/journal.pone.0026077].
41. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM: *Mixed Effects Models and Extensions in Ecology with R.* New York: Springer 2009.
42. Gelman A, Hill J: *Data Analysis using Regression and Multilevel/Hierarchical Models.* Cambridge, UK: Cambridge University Press 2007.
43. Stodgell CJ, Ingram JL, O'Bara M, Tisdale BK, Nau H, Rodier PM: **Induction of the homeotic gene Hoxal through valproic acid's teratogenic mechanism of action.** *Neurotoxicol Teratol* 2006, **28**(5):617–624, [http://dx.doi.org/10.1016/j.ntt.2006.08.004].
44. Kuwagata M, Ogawa T, Shioda S, Nagata T: **Observation of fetal brain in a rat valproate-induced autism model: a developmental neurotoxicity study.** *Int J Dev Neurosci* 2009, **27**(4):399–405, [http://dx.doi.org/10.1016/j.ijdevneu.2009.01.006].
45. Murawski NJ, Brown KL, Stanton ME: **Interstimulus interval (ISI) discrimination of the conditioned eyeblink response in a rodent model of autism.** *Behav Brain Res* 2009, **196**(2):297–303, [http://dx.doi.org/10.1016/j.bbr.2008.09.020].
46. Rodier PM, Ingram JL, Tisdale B, Nelson S, Romano J: **Embryological origin for autism: developmental anomalies of the cranial nerve motor nuclei.** *J Comp Neurol* 1996, **370**(2):247–261, [http://dx.doi.org/10.1002/jcompneu.10002].
47. Ingram JL, Peckham SM, Tisdale B, Rodier PM: **Prenatal exposure of rats to valproic acid reproduces the cerebellar anomalies associated with autism.** *Neurotoxicol Teratol* 2000, **22**(3):319–324.
48. Flood ZC, Engel DLJ, Simon CC, Negherbon KR, Murphy LJ, Tamavimok W, Anderson GM, Janusonis S: **Brain growth trajectories in mouse strains with central and peripheral serotonin differences: relevance to autism models.** *Neuroscience* 2012, **210**:286–295, [http://dx.doi.org/10.1016/j.neuroscience.2012.03.010].
49. Peers IS, Ceuppens PR, Harbron C: **In search of preclinical robustness.** *Nat Rev Drug Discov* 2012, **11**(10):733–734, [http://dx.doi.org/10.1038/nrd3849].
50. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, Crystal RG, Darnell RB, Ferrante RJ, Fillit H, Finkelstein R, Fisher M, Gendelman HE, Golub RM, Goudreau JL, Gross RA, Gubitza AK, Hesterlee SE, Howells DW, Huguenard J, Kelner K, Koroshetz W, Krainc D, Lazic SE, Levine MS, Macleod MR, McCall JM, Moxley RT, Narasimhan K, Noble LJ, Perrin S, Porter JD, Steward O, Unger E, Utz U, Silberberg SD: **A call for transparent reporting to optimize the predictive value of preclinical research.** *Nature* 2012, **490**(7419):187–191, [http://dx.doi.org/10.1038/nature11556].
51. Hooijmans CR, Leenaars M, Ritskes-Hoitinga M: **A gold standard publication checklist to improve the quality of animal studies, to fully integrate the Three Rs, and to make systematic reviews more feasible.** *Altern Lab Anim* 2010, **38**(2):167–182.
52. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG: **Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research.** *PLoS Biol* 2010, **8**(6):e1000412, [http://dx.doi.org/10.1371/journal.pbio.1000412].
53. Wicherts JM, Borsboom D, Kats J, Molenaar D: **The poor availability of psychological research data for reanalysis.** *Am Psychol* 2006, **61**(7):726–728, [http://dx.doi.org/10.1037/0003-066X.61.7.726].
54. Bakker M, Wicherts JM: **The (mis)reporting of statistical results in psychology journals.** *Behav Res Methods* 2011, **43**(3):666–678, [http://dx.doi.org/10.3758/s13428-011-0089-5].
55. Wicherts JM, Bakker M, Molenaar D: **Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results.** *PLoS One* 2011, **6**(11):e26828, [http://dx.doi.org/10.1371/journal.pone.0026828].

Figures

Figure 1 - Defining the experimental unit

Pregnant females are the experimental units because they are randomised to the treatment (e.g. valproic acid) or control conditions, and therefore $n = 6$ in this example. The three offspring within a litter will often be more alike than offspring from different litters ($\frac{\text{Between-litter variation}}{\text{Within-litter variation}} > 1$), and multiple offspring within a litter can be thought of as subsamples or “technical replicates”, even though these are the scientific unit of interest. Only the mean of the within-litter values are important when comparing treated and control groups. Using all of the offspring without averaging will result in an inflated sample size (pseudoreplication) when using standard analyses. Instead of averaging, one could randomly select only one animal from each litter, or use a nested or hierarchical model to appropriately partition the different sources of variation. The only way to increase sample size, and thus power, is to increase the number of litters used.

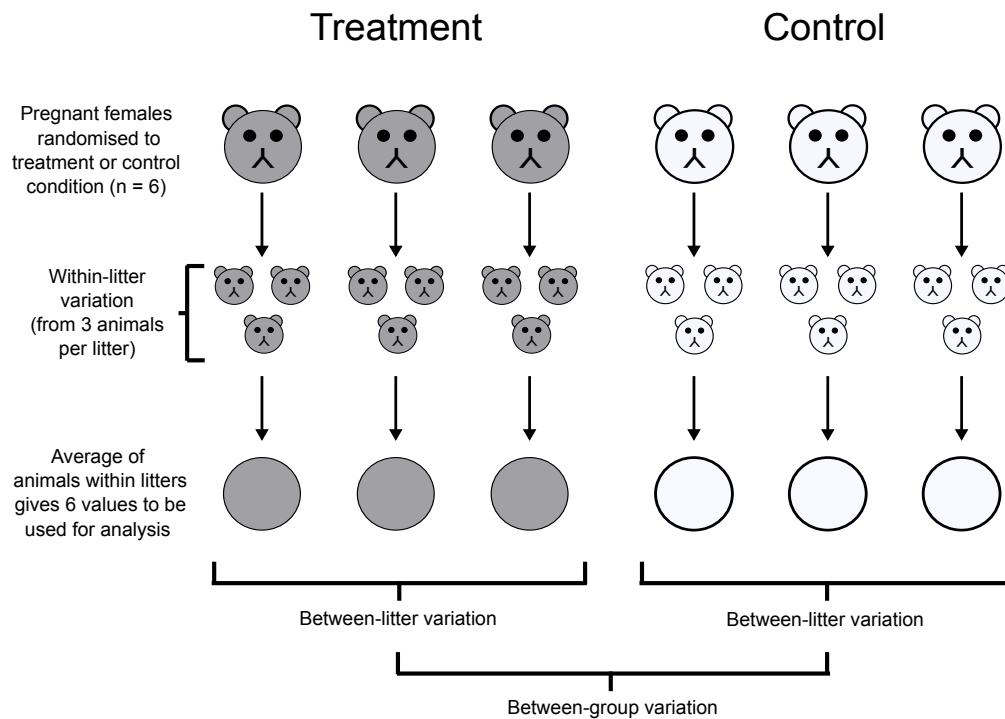


Figure 2 - Analysis with and without litter taken into account

Nine pregnant female C57BL/6 mice were injected with 600 mg/kg VPA subcutaneously on E13, and five control females received vehicle injections. Half of the animals in each condition were also injected with either a mGluR5 receptor antagonist (MPEP) or saline. Total locomotor activity in the open field over a 30 min period at 8–9 weeks of age is shown. There is a slight increase in activity due to MPEP, but this was not significant when differences between litters were ignored (Two-way ANOVA: mean difference = 0.60, $F(1,44) = 3.17$, $p = 0.082$). Adjusting for litter removed unexplained variation in the data, allowing the small difference between groups to become statistically significant (Hierarchical model: mean difference = 0.64, $F(1,32) = 7.19$, $p = 0.011$). Note how the values in the second graph have less variability around the group means; this increased precision increases the power of the statistical tests. Lines go through the mean of each group, and points are jittered in the x direction.

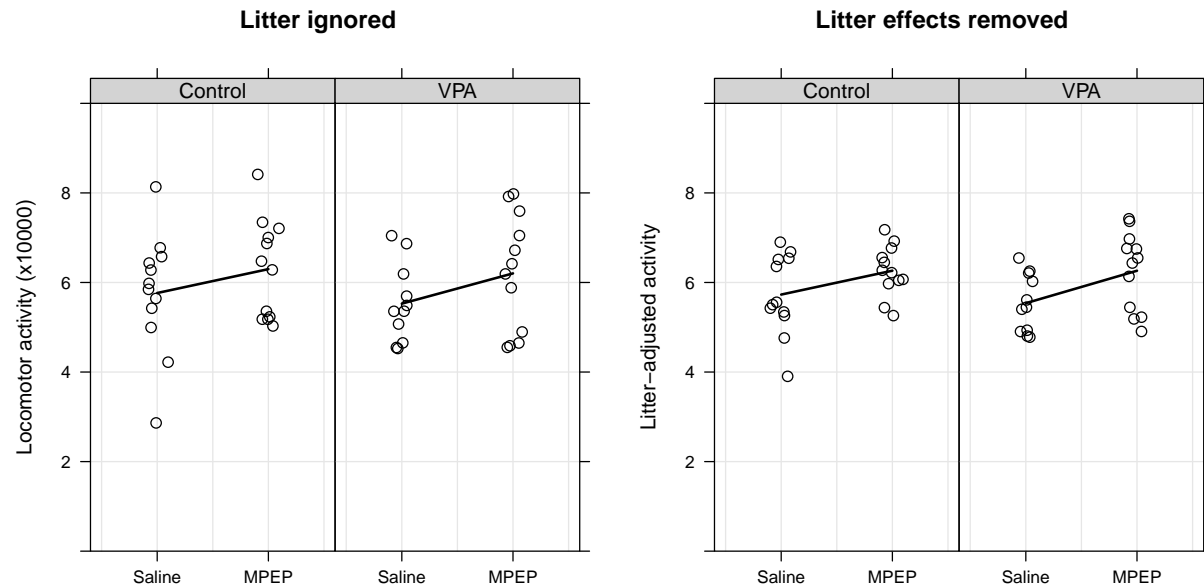


Figure 3 - Visualising litter-to-litter variation

The residuals represent the unexplained variation in the data after the effects of VPA and MPEP have been taken into account; they should be centred at zero and should not be associated with any other variable. However, the standard analysis (A) shows that when residuals are plotted against litter (x -axis) there are large differences between the different litters. In other words, there is another factor affecting the outcome besides the experimental factors of interest. The variance of the residuals (grey points on the right) is high ($\sigma_e^2 = 1.29$). The proper analysis (B) reduces the unexplained variation in the data by 61% ($\sigma_e^2 = 0.50$; $p < 0.001$), which can be seen by the narrower spread of the grey points around zero, and the large differences between the litters has been eliminated. This reduction in “noise” allows smaller true “signals” to be detected. Error bars are SEM. Litters F and L only have one observation and thus no error bars.

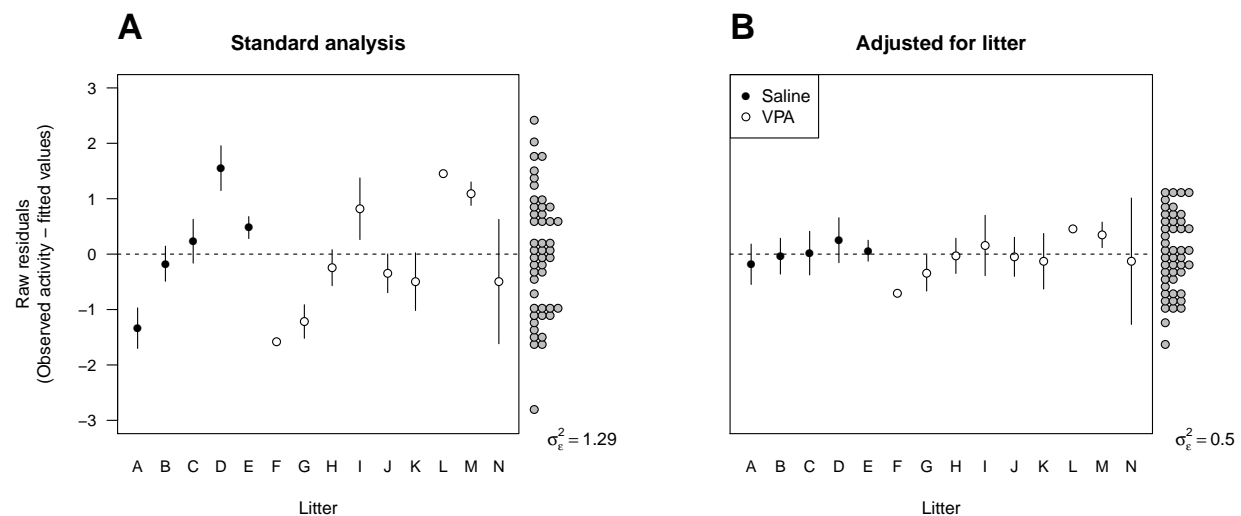
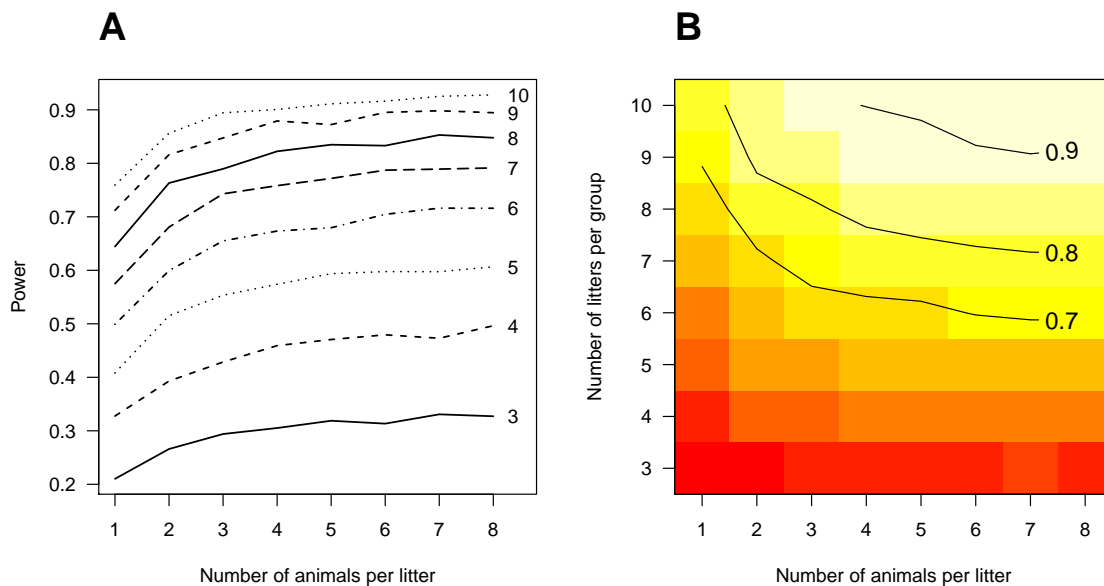


Figure 4 - Power calculations for VPA experiments

Panel A shows how power changes as the number of animals per litter increases from one to eight (x -axis) and the number of litters per group increases from three to ten (different lines). It is clear that increasing the number of animals per litter has only a modest effect on power, with little improvement after two animals. A two-group study with three litters per group and eight animals per litter ($2 \times 3 \times 8 = 48$ animals) will have only at 30% chance of detecting the effect. Whereas a study with ten litters per group and one animal per litter ($2 \times 10 \times 1 = 20$ animals) will have almost 80% power and use far fewer animals. Panel B shows the same data, but presented differently. Power for different combinations of litters and animals per litter is indicated by color (red = low power, white = high) and reference lines for 70%, 80% and 90% power are indicated. Note that these specific power values are only relevant for the locomotor activity task with a fixed effect size, and will have to be recalculated for other outcomes. However the general result (increasing litters is better than increasing the number of animals per litter) will apply for all outcomes.



Tables

Table 1 - Importance of litter effects on body weight and behavioural tests.

The p-value tests whether the litter-to-litter variation was significantly greater than zero.

Variable	Reduction in σ_ϵ^2	P-value
Locomotor activity	61%	<0.001
Body weight	50%	0.003
Marbles buried	38%	0.045
Anxiety (open field)	35%	0.0504
Grooming	23%	0.116

σ_ϵ^2 is the residual (unexplained) variation.

Additional Files

Additional file 1 — R code for the analyses and power calculations

Code for the analyses and power calculations are given as a plain text file.

Additional file 2 — Raw data

Raw data from Mehta et al. [40], including body weight, locomotor activity and anxiety measures from the open field test, grooming behaviour, and number of marbles buried in the marble-burying test. Details can be found in the original publication.

Additional file 3 — List of VPA studies

List of the thirty-four studies using the VPA rodent model of autism.

Additional file 4 — Power analysis for the mixed-effects model and the incorrect analysis

The interpretation of the graphs is the same as Figure 4 (main text). Panels A and B are for the mixed-effects model and are nearly identical to the results for averaging the values within each litter and then using a t-test (Figure 4 main text). Panels C and D ignore litter and just compare all of the data with a t-test, which results in an artificially inflated sample size and inappropriately high power.